GntR family of bacterial transcription factors and their binding motifs: structure and co-evolution

I.A. Suvorova

Institute for Information Transmission Problems RAS (the Kharkevich Institute), Bolshoi Karetny per. 19-1, Moscow, Russia, inn1313@yandex.ru

GntR family of transcription factors is a large group of proteins distributed among diverse bacteria. Here we use the comparative genomics approach to reconstruct regulons and identify binding motifs of the regulators from the three subfamilies of the GntR family: FadR, HutC, and YtrA. We report the results of the correlation analysis of the DNA binding sites and amino acid sequences of the HTH domains of transcription factors and predict the most favorable contacts.

Most of GntR-family binding sites are A/T-rich palindromic sequences with conserved GT/AC groups. We have shown that for FadR, HutC, and YtrA subfamilies nucleotide and amino acid positions that likely determine binding specificity correspond well to those identified for the protein-DNA structure of FadR (*E.coli*) [1].

The common consensus of all analyzed binding sites of FadR-subfamily TFs is taAAyTkGTm(t/-)kACmArTTta. Amino acids in position 28 of the HTH domain (here and further numeration starts from zero) show correlations with nucleotides in position 6 of the motif. Nucleotides in positions 6/14 are also correlated with amino acid residues in positions 40 and 59. Statistically significant predicted nucleotide-amino acid pairs likely include Arg-G, which is the most common pair, and also Gln-A, Pro-T, Ala-T, Asp-C, Ser-G, His-A, Asn-A, Gly-G. Though Gly-G might be not a direct protein-DNA contact, in FadR from *E.coli* glycine, occupying the same position, does not form specific contacts, but allows interaction of the nearby amino acid with DNA due to the lack a side chain [1]. Predicted unfavorable contacts are Arg-T, Gln-G, Ala-G, Ser-G, Pro-G, Gly-T, His-C, Asn-G.

The consensus sequence of all analyzed binding motifs of HutC-subfamily TFs, tataaAyTkGTmTAkACmArTttata, is very similar to the one of the FadR subfamily. Nucleotides 8/17 of the binding motif correlate with amino acid residues in positions 28 and

43 of the HTH domain. Predicted favorable nucleotide-amino acid pairs are Arg-G, Asn-A, Gly-G, Pro-G, Gln-C, Glu-C. Unfavorable pairs are likely Arg-T, Asn-G, Gln-T, Glu-T.

YtrA subfamily, its binding motifs and regulons have many features different from other studied subfamilies, e.g., binding motifs of TFs from the YtrA subfamily are significantly longer than motifs of other GntR-family TFs. Still, due to the conserved HTH domain structure in the GntR family, YtrA-type DNA-binding domains can be aligned accurately with domains from the other subfamilies. Correlations show that nucleotides 12/30 specifically interact with amino acids in position 28. Correlations are also observed for nucleotides 16-17/25-26 and amino acids in position 37. Predicted favorable nucleotide-amino acid pairs are Arg-G, Ser-A, Ser-T, Asn-A, Thr-A, Tyr-A.

It has been shown that amino acids that more frequently interact with nucleotides are Arg, Asn, Lys, Gln, Thr, Ser, Asp and Gly [2]. In general, hydrogen-bond donors residues (Arg, Cys, His, Lys, Ser, Thr) prefer G, while hydrogen bond acceptor residues (such as acidic Asp, Glu) prefer C, and Asn and Gln, possessing both donor and acceptor moieties, prefer A [2]. Our correlation data shows that most of predicted contacts in all analyzed subfamilies of the GntR family conform to the general interaction trends described in literature [2].

1. Xu Y. et al. (2001) The FadR-DNA Complex. Transcriptional control of fatty acid metabolism in *Escherichia coli*. *J Biol Chem* **276(20)**: 17373–17379

2. Marabotti A. et al. (2008) Energy-based prediction of amino acid-nucleotide base recognition. *J Comput Chem*. **29(12)**:1955-69